CS4400 Project

Chang Li

Code

The code for this solution is at: https://github.com/cli681/CS4400

1. Data reading

I first read all three tables including the left table, right table, as well as the training set. Next, to understand the data set better, I print out the number of entries for each column. From the data printed, I can see that the left table has 2554 and the right table has 22074 rows of entries. However, not all data is entered in attributes such as category, brand, modelno and price in the left table. Thus, when comparing values later, I should be mindful of these null values.

2. Blocking

Out of all the attributes, it makes the most sense to block by “brand”. And since I am finding pairs that match, it will not be meaningful if I found a brand in the right table that is not in the left table. Therefore, instead of taking the union of the brands of the two tables, I will only perform blocking based on the brands in the left table since it has way less entries. In addition, I decided to also block by “category” after blocking by brand. As two products with the same brand and same category are more likely to be the same entity. My blocking method reduces the number of pairs from 56376996 to 38662.

3. Feature engineering

I compute Jaro string similarity metric, similarity ratio and levenshtein distance for each pair of the candidate set to obtain a feature matrix. Having multiple similarity computation should increase the accuracy of the resulting pairs.

4. Model training

Since blocking cut down the number of pairs to a great level, I decided to go easy with the model training process. Therefore, using random forest classifier, I set several parameters to be the minimum in order to preserve most of the resulting pairs.

5. Generating output

All the pairs with a matching value of 1 are kept in the final resulting pairs. An csv file named “output” is generated to show the data.